Combining statistical machine translation and translation memories with domain adaptation

Läubli, Samuel ; Fishel, Mark ; Volk, Martin ; Weibel, Manuela

Abstract: Since the emergence of translation memory software, translation companies and freelance translators have been accumulating translated text for various languages and domains. This data has the potential of being used for training domain-specific machine translation systems for corporate or even personal use. But while the resulting systems usually perform well in translating domain-specific language, their out-of-domain vocabulary coverage is often insufficient due to the limited size of the translation memories. In this paper, we demonstrate that small in-domain translation memories can be successfully complemented with freely available general-domain parallel corpora such that (a) the number of out-of-vocabulary words (OOV) is reduced while (b) the in-domain terminology is preserved. In our experiments, a German–French and a German–Italian statistical machine translation system geared to marketing texts of the automobile industry has been significantly improved using Europarl and OpenSubtitles data, both in terms of automatic evaluation metrics and human judgement.

Originally published at:
Läubli, Samuel; Fishel, Mark; Volk, Martin; Weibel, Manuela (2013). Combining statistical machine translation and translation memories with domain adaptation. In: NODALIDA 2013, Nordic Conference of Computational Linguistics, Oslo, Norway, 22 May 2013 - 24 May 2013. Linköpings universitet Electronic Press, 331-341.

Combining Statistical Machine Translation and Translation Memories with Domain Adaptation

Samuel Läubli[1], *Mark Fishel*[1], *Martin Volk*[1], *Manuela Weibel*[2]

(1) Institute of Computational Linguistics, University of Zurich, Binzmühlestrasse 14, CH-8050 Zürich
(2) SemioticTransfer AG, Bruggerstrasse 37, CH-5400 Baden

{laeubli,fishel,volk}@cl.uzh.ch manuela.weibel@semiotictransfer.ch

ABSTRACT

Since the emergence of translation memory software, translation companies and freelance translators have been accumulating translated text for various languages and domains. This data has the potential of being used for training domain-specific machine translation systems for corporate or even personal use. But while the resulting systems usually perform well in translating domain-specific language, their out-of-domain vocabulary coverage is often insufficient due to the limited size of the translation memories. In this paper, we demonstrate that small in-domain translation memories can be successfully complemented with freely available general-domain parallel corpora such that (a) the number of out-of-vocabulary words (OOV) is reduced while (b) the in-domain terminology is preserved. In our experiments, a German–French and a German–Italian statistical machine translation system geared to marketing texts of the automobile industry has been significantly improved using Europarl and OpenSubtitles data, both in terms of automatic evaluation metrics and human judgement.

KEYWORDS: Machine Translation, Translation Memory, Domain Adaptation, Perplexity Minimization.

1 Introduction

Technological advances in the field of automated translation have increased economic pressure on translation companies and freelance translators. According to a recent study published by the European Union (Pym et al., 2012), there is evidence that the per-word rate for professional translations has decreased significantly in some western European countries.

On the other hand, computer-assisted translation tools such as translation memory (TM) systems—allowing translators to store their translations in a personal or corporate database and reuse them on future occasions—have become an integral part of state-of-the-art translation workflows, even for freelancers.

More recently, there have been efforts to combine machine translation systems with translation memories (Kanavos and Kartsaklis, 2010; Koehn and Senellart, 2010). As a result, major commercial systems such as SDL TradosStudio[1], Across LanguageServer[2], and even open source alternatives such as OmegaT[3] now offer machine translation interfaces, allowing translators to have segments automatically (pre-)translated in case there is no corresponding translation available in their translation memory.

In a joint project between the UZH Institute of Computational Linguistics and SemioticTransfer AG, we currently explore the potential of using domain-specific statistical machine translation (SMT) systems in human-based translation workflows. We hypothesize that SMT systems trained on SemioticTransfer's translation memory data will enable their translators to work more efficiently while preserving translation quality.

In our present paper, we report on training and evaluating statistical machine translation systems for two language pairs on a relatively small amount of marketing texts related to the automobile industry. In Section 2, we outline the foreseen translation scenario and introduce means of compensating for the limited size of our in-domain training material. In Section 3, we present SMT systems that are based on a combination of in-domain and out-of-domain data. The combined systems are compared against an in-domain-only baseline. We discuss our findings in Section 4, and lastly, we conclude and outline the further course of our project in Section 5.

2 Background

In this section, we first describe the translation scenario that our domain-specific SMT systems should be applied in. Second, we detail techniques for combining limited amounts of in-domain with out-of-domain translation data.

2.1 Translation Scenario

In SemioticTransfer's current translation workflow, new language material to be translated is imported into a corporate translation memory system by specially trained staff termed translation managers. Subsequently, the manager prepares one or several work packages, each of which consists of a number of documents in the source language with an accompanying empty document in the target language. Next, each source language segment is automatically looked up in the translation memory, and exact matches are directly inserted into the target

[1] http://www.sdl.com/products/sdl-trados-studio/index-tab3.html
[2] http://www.across.net/en/across-machine-translation-integration.aspx
[3] http://www.omegat.org/en/howtos/google_translate.html

language document. A specialist translator then translates all remaining segments one-by-one. This process is again supported by the translation memory system: For each segment, the translator is shown similar segments that are already stored in the TM—it is her or his decision whether to take an existing segment and adapt it or translate the segment from scratch instead.

There are many ways to integrate machine translation in such a workflow. Kanavos and Kartsaklis (2010) propose a pragmatic approach that uses a simple SMT model trained on all available translation memory data. This model is used to translate segments for which no fuzzy match with more than 80% similarity is available in the TM, which are subsequently post-edited by specialist translators. The automatic translation is either invoked in an "on-demand" (the translator requests an SMT hypothesis for a segment he is currently translating) or a "one time" mode. We focus on the latter which automatically inserts SMT translations for segments with fuzzy matches below 80% into the target language document before it is handed to the translator, i.e., we produce a pre-translation which is then completed and post-edited inside the translation memory system.

2.2 Domain Adaptation

Unlike Kanavos and Kartsaklis (2010), we focus on a specific text genre: marketing texts of the automobile industry. For this reason, we cannot use large general-domain parallel corpora for training our translation models as (a) many domain-specific terms are not contained in such resources and (b) domain-specific terms would be prone to mistranslation because of domain mismatches. For example, the German word *Arten* could be translated as *espèces* (species) in a general-domain DE–FR translation system. While this translation could be used, for example, in biological texts, it is not adequate in the automobile domain as cars are of certain *types* (types) rather than species.

To counteract these problems, Koehn and Senellart (2010) have proposed to retrieve a fuzzy match in the TM for each source segment to be translated, identify the mismatched parts, and replace these parts by an SMT translation. Their approach relies on automatic word alignment to find the target words that are affected by the mismatch.

In contrast to Koehn and Senellart (2010) we approach combining parallel texts from domain-specific translation memories and general-domain corpora as a domain adaptation problem. We use the approach of mixture-modeling, commonly used for language model adaptation and extended to translation models by Foster and Kuhn (2007). The main distinctive feature of this approach for both language and translation models is that instead of separating the data and models into in-domain/out-of-domain in a binary setting, different domains are assigned real-valued weights, reflecting their similarity to in-domain text material. Using these weights the single domain models $\mathbf{p} = \{p_i\}_{i=1\ldots N}$ are combined into a single adapted model:

$$p(\mathbf{x}) = \mathbf{w}\mathbf{p}(\mathbf{x}) = \sum_{i=1}^{N} w_i p_i(\mathbf{x}).$$

The weights $\mathbf{w}$ are selected to optimize the performance of the adapted model on an in-domain development set. Language model performance is estimated with its entropy on the said set. Unlike Foster and Kuhn (2007) who use a monolingual performance measure for translation

models we follow Sennrich (2012) and use the cross-entropy of the adapted translation model on the development set, a bilingual performance measure:

$$\hat{\mathbf{p}}(\mathbf{x}) = \arg\min_{\mathbf{w}} H(p), \text{ where } H(p) = -\sum_{\mathbf{x} \in \mathbf{X}^{(d)}} \tilde{p}(\mathbf{x}) \log_2 p(\mathbf{x})$$

Here $H(p)$ is the cross-entropy of the adapted language or translation model p and $\mathbf{X}^{(d)}$ is the development set. The empirical probability distribution $\tilde{p}$ is based on the development set. The nature of $\mathbf{x} \in \mathbf{X}^{(d)}$ depends on which models are handled: in case of language models, it represents a single sentence; in case of translation models it stands for a $\langle source, translation \rangle$ tuple. For more details see (Sennrich, 2012; Chen and Goodman, 1998).

3 Combining In-Domain with Out-of-Domain Data

In this section, we describe the statistical machine translation systems we have trained for our experiments. We briefly describe the training data (Section 3.1) and the system setups (Section 3.2), which we have evaluated by automatic means (Section 3.3.1) as well as with a human evaluator (Section 3.3.2).

3.1 Training Data

3.1.1 In-Domain

The in-domain data has been extracted from SemioticTransfer's translation memories for German–French (DE–FR) and German–Italian (DE–IT). All texts are related to the automobile industry; the translated segments stem from brochures, websites, and price lists. For this reason, a segment may refer to a whole sentence, but also to smaller units such as short phrases or even single words of table entries. After cleaning the memories, we were able to extract 166'957 segments for DE–FR and 112'166 segments for DE–IT, which corresponds to ∼2.0 and ∼1.5 million tokens, respectively. Please note that all segments are unique. Unlike text from a corpus of running words, each segment from a translation memory occurs only once. The exact numbers are shown in Table 1.

3.1.2 Out-of-Domain

As for out-of-domain data, we have chosen two freely available parallel corpora: Europarl v7 (EP7) (Koehn, 2005) and OpenSubtitles 2011 (OS11) (Tiedemann, 2009). We extracted the DE–FR and DE–IT translations from each of them, resulting in ∼48.5 (EP7) and∼16.0 (OS11) million tokens per language pair. See Table 1. We point out that these parallel corpora are not thematically related to our in-domain data. Rather than that, they are much more extensive and thus cover a broader vocabulary, which is missing in our in-domain data due to its limited size.

	In-Domain		Europarl		OpenSubtitles	
	DE–FR	DE–IT	DE–FR	DE–IT	DE–FR	DE–IT
Segments	166'957	112'166	1'903'628	1'805'792	2'852'474	2'131'004
Tokens F	2'011'872	1'413'452	48'405'406	48'419'389	16'858'070	15'642'379
Tokens E	2'632'256	1'731'219	56'372'702	50'689'987	16'370'845	15'458'666
Tokens/Segment F	12.05	12.60	25.43	26.81	5.91	7.34
Tokens/Segment E	15.77	15.43	29.61	28.07	5.74	7.25

Table 1: In-Domain and Out-of-Domain Training Data

3.2 Domain-Specific SMT Systems

We have trained an in-domain-only baseline and three systems that combine in-domain and out-of-domain data (see Section 3.1) in different weighting modes for each language pair.

3.2.1 Baseline Systems

Using only the in-domain data, we trained a standard phrase-based statistical machine translation system for DE–FR and DE–IT. 5-gram language models were trained using the IRST Language Modeling Toolkit (Federico and Cettolo, 2007). Otherwise, we relied on the Moses decoder (Koehn et al., 2007) and its dedicated scripts for training, tokenization, and truecasing.

3.2.2 Combined Systems

Additionally, we trained separate translation and language models using our out-of-domain corpora. Apart from the training data, the setup was the same as for our baseline systems. This left us with three phrase tables and language models per language pair: in-domain, EP7, and OS11. All combined systems presented in this section comprise all of these models, but they are differently weighted. The weighting modes are defined as follows.

Unweighted Combination (*unweighted*)

In the *unweighted* mode, we simply concatenated the in-domain, EP7, and OS11 language material. On this basis, we trained a combined translation and language model for each language pair.

Weighted Language Model (*weighted LM*)

In the *weighted LM* mode, we used an interpolated language model alongside an *unweighted* translation model for each language pair. We employed `interpolate-lm` of the IRST Language Modeling Toolkit (Federico and Cettolo, 2007) to estimate interpolation weights for the in-domain, EP7, and OS 11 language models by minimizing their perplexity on a development set of 2'000 in-domain segments (see section 2.2).

	DE–FR		DE–IT	
Corpus	**TM**	**LM**	**TM**	**LM**
In-Domain	-	91.9	-	92.9
EP7	-	5.8	-	5.0
OS11	-	2.3	-	2.1

Interpolation weights for translation (TM) and language (LM) models in %, *weighted LM* mode.

Weighted Translation Model (*weighted TM*)

	DE–FR		DE–IT	
Corpus	**TM**	**LM**	**TM**	**LM**
In-Domain	92.3	-	93.7	-
EP7	6.2	-	5.1	-
OS11	1.5	-	1.2	-

Interpolation weights for translation (TM) and language (LM) models in %, *weighted TM* mode.

In the *weighted TM* mode, we combined an interpolated translation model with an *unweighted* language model for each language pair. We applied Sennrich's approach (Sennrich, 2012) for estimating interpolation weights for the in-domain, EP7, and OS11 translation models, again using a development set of 2'000 in-domain segments (see section 2.2).

Weighted Language and Translation Model (*weighted LM+TM*)

In the *weighted LM+TM* mode, we used *both* interpolated language and translation models, i.e., we combined the interpolated language models of the *weighted LM* mode with the interpolated translation models of the *weighted TM* mode for each language pair.

	DE–FR		DE–IT	
Corpus	**TM**	**LM**	**TM**	**LM**
In-Domain	92.3	91.9	93.7	92.9
EP7	6.2	5.8	5.1	5.0
OS11	1.5	2.3	1.2	2.1

Interpolation weights for translation (TM) and language (LM) models in %, *weighted LM+TM* mode.

3.3 Evaluation

3.3.1 Automatic Evaluation

For evaluating the systems described in Section 3.2, we compiled a test set of 500 segments for both language pairs. The segments were taken from a recent contract of SemioticTransfer: the translation of an automobile company's website from German to French and Italian. The translation was carried out by two professional translators using SemioticTransfer's translation memories.

We used `multeval` (Clark et al., 2011) for evaluating our systems. As there is no full METEOR support for Italian (Denkowski and Lavie, 2011), this score is only available for the DE–FR system. Table 2 gives an overview on all metrics and the performance of the systems; a discussion of these and other results is given in Section 4.

		DE–FR				DE–IT			
Metric	**Mode**	**Avg**	$\bar{s}_{\text{sel}}$	s_{Test}	p-**value**	**Avg**	$\bar{s}_{\text{sel}}$	s_{Test}	p-**value**
BLEU ↑	baseline	32.2	1.3	0.1	-	31.3	1.4	0.3	-
	unweighted	31.3	1.3	0.1	0.00	30.8	1.4	0.2	0.12
	weighted LM	31.8	1.3	0.2	0.15	32.9	1.5	0.3	0.00
	weighted TM	32.8	1.3	0.3	0.00	31.9	1.4	0.0	0.02
	weighted LM+TM	33.0	1.3	0.2	0.00	31.9	1.4	0.1	0.00
TER ↓	baseline	51.7	1.2	0.3	-	55.9	1.4	0.6	-
	unweighted	52.9	1.2	0.1	0.00	54.3	1.3	0.5	0.00
	weighted LM	51.7	1.2	0.4	0.99	51.9	1.3	0.6	0.00
	weighted TM	50.9	1.2	0.6	0.00	54.7	1.4	0.0	0.00
	weighted LM+TM	50.6	1.2	0.6	0.00	53.3	1.3	0.5	0.00
METEOR ↑	baseline	50.5	1.1	0.1	-				
	unweighted	50.1	1.1	0.2	0.12				
	weighted LM	50.8	1.1	0.2	0.10				
	weighted TM	51.6	1.1	0.3	0.00				
	weighted LM+TM	51.7	1.1	0.3	0.00				

Table 2: Automatic Evaluation. Baseline and Combined Systems. p-values are relative to baseline and indicate whether a difference of this magnitude (between the baseline and the system on that line) is likely to be generated again by some random process (a randomized optimizer). Metric scores are averages over 5 MERT runs (Och, 2003; Bertoldi et al., 2009). s_{sel} indicates the variance due to test set selection and has nothing to do with optimizer instability.

	DE–FR			DE–IT		
Mode	**baseline**	**eq.**	**weighted TM**	**baseline**	**eq.**	**weighted TM**
general	15.3	48.7	36.0	15.3	30.7	54.0
terminology	14.7	48.7	36.7	15.3	30.7	54.0

Table 3: Human Evaluation. Subjective preference for 150 Segments translated by the *baseline* and *weighted TM* systems in terms of general and terminological accuracy in %. "eq." indicates that both translation hypotheses were considered equally well.

3.3.2 Human Evaluation

In order to reinforce our automatic evaluation, we conducted a further experiment with a human evaluator who is familiar with SemioticTransfer's automobile domain terminology. We wanted to assess whether and to which extent the difference between the baseline's and a combined system's scores was perceptible for an actual human translator. We chose the *weighted TM* systems for the comparison as preliminary results suggested negligible differences between *weighted TM* and *weighted LM+TM* scores. At the time, we concluded that weighting the language model in addition to the translation model is not worth the additional effort.

We provided our subject with a source segment (DE) alongside a reference translation (FR/IT) and two translation hypotheses for 150 randomly selected segments. One hypothesis was produced by the *baseline*, the other by the *weighted TM* system. The subject's task was to rate which hypothesis was better, (a) in general and (b) with regard to domain technology, with ties allowed. This evaluation setup, known as pairwise system comparison, has lately been favored by the MT community for it is simpler and better reproducible than, e.g., fluency/adequacy judgements on a five-point scale (Callison-Burch et al., 2012). Furthermore, genuine differences between two systems can easily be quantified using the Sign Test for paired observations.

The results, shown in Table 3, reveal the evaluator's clear preference towards the *weighted TM* system's translations. This holds especially for DE–IT. Statistically, the *weighted TM* systems outperform the *baseline* in all regards at $p \leq 0.001$, except for the adequacy of domain terminology in the DE–FR systems ($p \leq 0.01$).

4 Discusson

Our evaluation (see Section 3.3) shows that adding large amounts of out-of-domain data without adequate weights is no suitable means of improving our in-domain-only baseline systems (see Table 2, *unweighted* mode). An unweighted combination with the out-of-domain models lowers the scores for DE–FR compared to the baseline. The difference is statistically significant for BLEU ($p \leq 0.01$) and TER ($p \leq 0.01$), but not for METEOR. For DE–IT, TER improves ($p \leq 0.01$), but at the same time, BLEU decreases slightly (though not significantly). Insights into corresponding translations reveal that the added material reduces the number of out-of-vocabulary (OOV) words, but this effect is by far outclassed by the number of domain-specific terms and phrases that get translated in an inadequate way in the *unweighted* mode (see Table 5).

The *weighted LM* mode has opposing effects on the two language pairs. On one hand, the language model adaptation highly affects DE–IT. *Weighted LM* performs best out of all modes for this language pair and outperforms the baseline significantly, both in terms of BLEU ($p \leq 0.01$) and TER ($p \leq 0.01$). This improvement is in line with the finding of Foster and Kuhn (2007) that language model adaptation works well. Conversely, an interpolated language model has no positive effect on the DE–FR language pair.

Set	Feature	DE–FR $H_{unweighted}$	DE–FR $H_{weighted}$		DE–IT $H_{unweighted}$	DE–IT $H_{weighted}$	
DEV	inverse phrase translation probability	2.21	2.18		1.90	1.88	
	inverse lexical weighting	6.91	5.85	–	7.19	5.75	– –
	direct phrase translation probability	2.64	2.59		2.14	2.06	
	direct lexical weighting	7.35	6.13	–	7.66	5.92	– –
TEST	inverse phrase translation probability	2.91	3.32	+	3.39	3.78	+
	inverse lexical weighting	5.24	5.22		5.53	5.55	
	direct phrase translation probability	3.63	3.88		3.37	3.77	+
	direct lexical weighting	6.38	5.81		5.82	5.49	

Table 4: Cross entropies H between the DEV and TEST sets and the *unweighted* and *weighted* phrase tables. Plus and minus signs denote changes of ≥10% (+/−) and ≥20% (++ / −−).

In contrast, weighting the translation models (*weighted TM* mode) leads to a significant improvement of *all* scores in *both* language pairs (all at $p \leq 0.01$, except $p \leq 0.05$ for BLEU in DE–IT). Here, it turns out that using Sennrich's approach (Sennrich, 2012) to estimate weights for in-domain and out-of-domain translation models is a promising way for improving our systems. The high weights given to the in-domain models ensure that translations in the domain-specific phrase tables are normally preferred over identical source segments with different translations in the out-of-domain data (see Table 5). At the same time, the added out-of-domain translations lower the OOV rate in the *weighted TM* systems, no matter how small their translation probabilities become due to the weighted combination. In our DE–FR test set, this corresponds to a reduction from 194 to 124 OOV types (-36.1%) .

Using both a weighted language and translation model results in no further improvement. For DE–FR, there is a small increase in BLEU and TER, as well as a slight decrease in METEOR, but neither of these changes is significant. For DE–IT however, adding the weighted translation models on top of the weighted language models significantly decreases BLEU and TER (both at $p \leq 0.01$). This can only happen if the development set is too far apart from the test set – in other words, the development set is a poor generalization over the in-domain material, as a result of which the weighting overfits to it. Comparison of the unoptimized and optimized cross-entropies (presented in Table 4) confirms this conclusion: the cross-entropies naturally decrease on the development set both in case of DE–FR and DE–IT as the result of optimization; however, optimization also leads to increased cross-entropy of the combined model on the DE–IT test set.

Further experiments are needed to clarify the influence of interpolated language and translation models, as well as their interplay, on translation quality. Our evaluation of the combined DE–FR systems stands in contrast to the claim by Foster and Kuhn (2007) that "LM adaptation works well, and adding an adapted TM yields no improvement"—rather than that, TM combination works well, and adding an adapted LM yields no significant improvement in our case. However, we cannot generalize this finding to the combined DE–IT systems.

Apart from that, it is remarkable that the estimated interpolation weights of all in-domain models clearly exceed 90%, both in the DE–FR and DE–IT systems (see Section 3.2.2). The high percentages confirm that the marketing texts we look at are very different from our out-of-domain language material, which justifies the development of domain-specific SMT models instead of using out-of-the-box systems that are mostly trained on general-domain resources.

f	$\arg\max_e \varphi(e\|f)_{unweighted}$	$\arg\max_e \varphi(e\|f)_{weighted}$
Abgasemissionen	émissions de gaz	émissions de gaz d'échappement
Arten	espèces	types
Decke	couverture	plafond
Kiste	boîte	caisse
Klappe	ferme	clapet
Linksverkehr	conduite	conduite à gauche
PKW	voitures	voiture particulière
Tempo	rythme	vitesse
Unterbodenverkleidung	caisse	garnitures du dessous de caisse
Werke	œuvres	usines
Zulassungen	autorisations	immatriculations
Zünder	détonateur	amorce

Table 5: French terms e that maximize the translation probabilities in the unweighted (*unweighted* mode) and weighted (*weighted TM* and *weighted LM+TM* modes) translation models of the DE–FR systems, given a selection of German terms f.

5 Conclusion and Future Work

In our present study on using domain-specific translation memories for training machine translation systems, we have shown that very limited amounts of in-domain data can be successfully complemented with general-domain parallel corpora for improving translation performance. Our experiments demonstrate that assigning adequate weights to the in-domain and out-of-domain language and translation models is crucial for successful combinations. Our evaluations confirm that while the number of OOV types decreases through adding large amounts of out-of-domain data, the in-domain terminology prevails over alternative translations due to the weighting. In our German–French and German–Italian systems, using both an interpolated language and translation model has resulted in significant BLEU, METEOR and TER increases. These performance gains have been confirmed by a human evaluator who is familiar with the domain terminology.

In the further course of our project, we would like to give particular attention to German compounds, which constitute approximately 65% of the remaining OOV words in our combined systems, depending on test set and target language. We would like to assess in a systematic way which of the numerous approaches to decompounding such as (Koehn and Knight, 2003; Dyer, 2009; Stymne, 2009; Hardmeier et al., 2010) is suitable for our data and translation scenario. Ultimately, we plan to conduct targeted *in-situ* experiments in order to measure the impact of using our domain-adapted SMT systems in their designated human-based translation workflows.

Acknowledgements

We would like to thank Rico Sennrich as well as our anonymous reviewers for their valuable feedback and support. This work was supported by Grant No. 11926.2 PFES-ES from the Swiss Federal Commission for Technology and Innovation CTI.

References

Bertoldi, N., Haddow, B., and Fouet, J.-B. (2009). Improved minimum error rate training in moses. *The Prague Bulletin of Mathematical Linguistics*, 91(1):7–16.

Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2012). Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada. Association for Computational Linguistics.

Chen, S. F. and Goodman, J. (1998). An empirical study of smoothing techniques for language modeling. *omputer Speech & Language*, 13:359–393.

Clark, J. H., Dyer, C., Lavie, A., and Smith, N. A. (2011). Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 176–181, Stroudsburg, PA, USA. Association for Computational Linguistics.

Denkowski, M. and Lavie, A. (2011). Meteor 1.3: automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, WMT '11, pages 85–91, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dyer, C. (2009). Using a maximum entropy model to build segmentation lattices for MT. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 406–414, Stroudsburg, PA, USA. Association for Computational Linguistics.

Federico, M. and Cettolo, M. (2007). Efficient handling of n-gram language models for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 88–95, Stroudsburg, PA, USA. Association for Computational Linguistics.

Foster, G. and Kuhn, R. (2007). Mixture-model adaptation for SMT. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 128–135, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hardmeier, C., Bisazza, A., and Federico, M. (2010). FBK at WMT 2010: word lattices for morphological reduction and chunk-based reordering. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pages 88–92, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kanavos, P. and Kartsaklis, D. (2010). Integrating machine translation with translation memory: A practical approach. In *JEC 2010: Second joint EM+/CNGL Workshop "Bringing MT to the user: research on integrating MT in the translation industry"*, pages 11–20.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Machine Translation Summit X*, pages 79–86, Phuket, Thailand.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Koehn, P. and Knight, K. (2003). Empirical methods for compound splitting. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, pages 187–193, Stroudsburg, PA, USA. Association for Computational Linguistics.

Koehn, P. and Senellart, J. (2010). Convergence of translation memory and statistical machine translation. In *JEC 2010: Second joint EM+/CNGL Workshop "Bringing MT to the user: research on integrating MT in the translation industry"*, pages 21–31.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 160–167, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pym, A., Grin, F., Sfreddo, C., and Chan, A. L. J. (2012). *The Status of the Translation Profession in the European Union*, volume 7/2012 of *Studies on translation and multilingualism*. Publications Office of the European Union.

Sennrich, R. (2012). Perplexity Minimization for Translation Model Domain Adaptation in Statistical Machine Translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 539–549, Stroudsburg, PA, USA. Association for Computational Linguistics.

Stymne, S. (2009). Compound processing for phrase-based statistical machine translation. Master's thesis, Linköping University, Sweden.

Tiedemann, J. (2009). News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In Nicolov, N., Bontcheva, K., Angelova, G., and Mitkov, R., editors, *Recent Advances in Natural Language Processing*, volume V, pages 237–248, Borovets, Bulgaria. John Benjamins, Amsterdam/Philadelphia.